

AVA: Towards Autonomous Visualization Agents through Visual Perception-Driven Decision-Making

Shusen Liu^{1*†}, Haichao Miao^{1*}, Zhimin Li², Matthew Olson¹, Valerio Pascucci² and Peer-Timo Bremer^{1,2}
¹Lawrence Livermore National Laboratory, ²SCI Institute, University of Utah

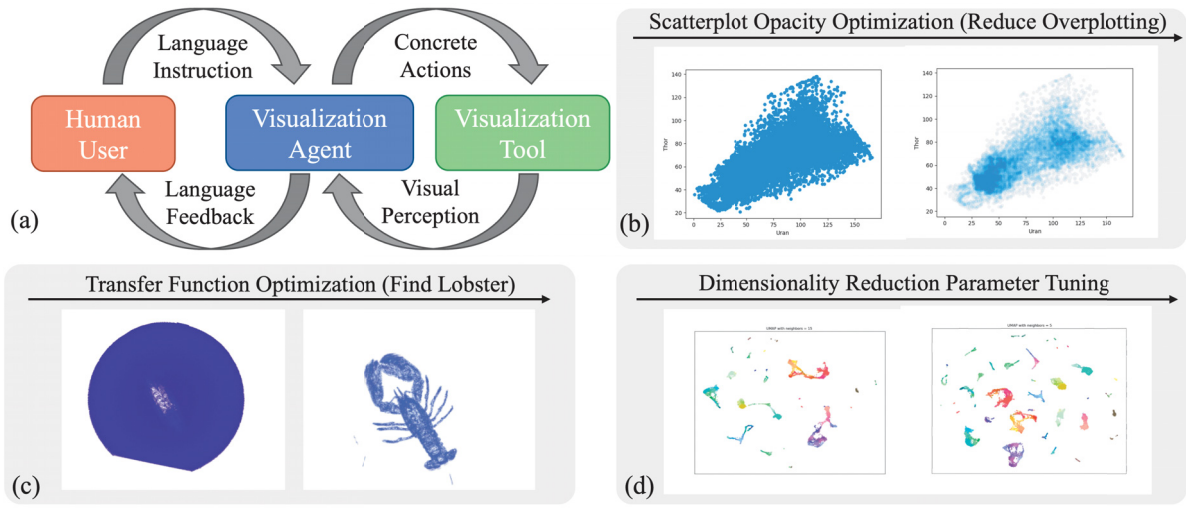


Figure 1: Overview of Autonomous Visualization Agents (AVAs): By combining natural language understanding with visual perception (a) AVAs can not only understand user instructions but also control and adjust a visualization system by interpreting its visual outputs to accomplish user-specified goals. We demonstrate the broad applicability of the proposed paradigm in multiple distinct scenarios including scatterplot opacity selection (b), volume rendering (c), and hyperparameter tuning for nonlinear dimensionality reduction (d).

Abstract

With recent advances in multi-modal foundation models, the previously text-only large language models (LLM) have evolved to incorporate visual input, opening up unprecedented opportunities for various applications in visualization. Our work explores the utilization of the visual perception ability of multi-modal LLMs to develop Autonomous Visualization Agents (AVAs) that can interpret and accomplish user-defined visualization objectives through natural language. We propose the first framework for the design of AVAs and present several usage scenarios intended to demonstrate the general applicability of the proposed paradigm. The addition of visual perception allows AVAs to act as the virtual visualization assistant for domain experts who may lack the knowledge or expertise in fine-tuning visualization outputs. Our preliminary exploration and proof-of-concept agents suggest that this approach can be widely applicable whenever the choices of appropriate visualization parameters require the interpretation of previous visual output. Feedback from unstructured interviews with experts in AI research, medical visualization, and radiology has been incorporated, highlighting the practicality and potential of AVAs. Our study indicates that AVAs represent a general paradigm for designing intelligent visualization systems that can achieve high-level visualization goals, which pave the way for developing expert-level visualization agents in the future.

1. Introduction

Within just a few months, large language models (LLMs) have been widely adapted to solve a variety of tasks [WXJ*23, WBZ*23,

* Contribute equally

† Project lead

SWW*23, WMF*23]. In the visualization domain, these models have been used to produce visualizations [DD19, Dib23] either through visual grammars like *Vegalit* [SMWH17], or directly generating visualization code (e.g., in Matplotlib [Hun07], VTK [SK19]). However, due to the inherently visual nature of these systems, purely language-based models have limited capability to make sense of their output. This significantly hampers or even prevents the analysis of the results and thus severely limits the opportunity for iterative interactions with the given visualization systems. The recent introduction of multimodal LLMs, such as GPT-4V, has the potential to address this fundamental limitation by filling the visual perception gap, which opens many possibilities for new paradigms of interaction between existing visualization tools and human users.

One particularly interesting and powerful usage is the adoption of an Autonomous Visualization Agent (AVA) that can act as the medium between domain experts and visualization tools to facilitate and enrich user interaction (see Figure 1(a)). Here, the AVA is defined as an entity that can understand high-level instructions (i.e., natural language) and autonomously carry out a sequence of actions in a visualization system based on the model’s prior knowledge. More specifically, given the ability to perceive the visualization output an AVA can adjust and refine the parameters to meet the initial user-specified goal. Such an agent will not only be able to relieve the user from potentially tedious and repetitive tasks but will also accomplish non-trivial visualization goals by iterative refining the existing visualization through visual feedback (following the *visualization* $\rightarrow$ *perception* $\rightarrow$ *action* paradigm, just as a visualization expert would do).

The agent-assisted visualization paradigm has the potential to fundamentally change how users interact with existing and future visualization tools. Despite enormous efforts from the visualization community to design user-friendly approaches, many standard visualization tools remain feature-rich and challenging to navigate for a wide range of users with diverse backgrounds. For example, many experts in the application domains (e.g. industrial or medical) continue finding that designing an effective transfer function [LKG*16] for volume rendering is a non-intuitive and challenging process [FAT99]. In other words, there often remains a fundamental knowledge gap between the developers who design the visualizations and the intended target users of such systems. Visualization agents tailored for each tool can act as virtual assistants that bridge this knowledge gap and enable non-visualization experts to easily control, steer, and iterate the visualization based on high-level objectives specified by natural language.

Here, we aim to take the first step towards making AVAs a reality by exploring their design space and demonstrating their capability for solving real-world visualization tasks across different visualization areas. The key power of AVAs derives from their ability to detect visual features associated with natural language instruction. Consequently, they can evaluate complex objectives that cannot be easily expressed algorithmically, i.e., *is there a particular structure in the rendering results?* or *does overplotting exist in the given scatterplot?* Despite its power, visual perception capabilities are only part of an agent. Once we obtain the visual understanding, the agent needs to plan its actions to achieve the goal. As illustrated in Figure

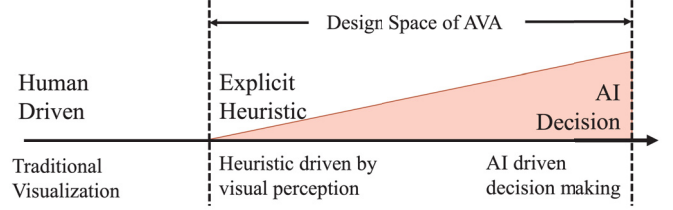


Figure 2: The design space of AVAs. On one end, we explicitly encode heuristics on how to update the visualization parameter, i.e., how a transfer function should be changed, which is driven by a high-level objective specified through language, i.e., "does this show the structure of interest". Alternatively, we can aim for a fully self-directed system with no explicit guidance on its action beyond the initial instruction (prompt).

2, this presents a range of possible designs. On one end (i.e., more explicit control), we can rely on heuristics to dictate the response. This is achieved by encoding our prior domain knowledge into decision rules. Alternatively, we can rely on the LLM and its prior knowledge to process the observations and plan the next action in a fully self-directed fashion.

To design an effective AVA, we first need to understand the capability and limitation of visual perception of the state-of-the-art multimodal LLMs (we use GPT4-Vision in all of our studies). We carried out a preliminary exploration of a few perception tasks related to common visualization outputs, including volume rendering, scatterplots, parallel coordinate plots, and graphs. Leveraging what we learned from these simple benchmarks, we avoid areas of visualization where the visual perception of the current models is performing less accurately (e.g., graph, parallel coordinate).

Since different visualization tasks require potentially vastly different knowledge and strategies, instead of designing a general agent for arbitrary tasks, we approach the problem by designing specialized AVAs for different use cases under a common base agent implementation. To demonstrate the feasibility and broad applicability of the proposed scheme, we intentionally select a distinctive set of applications, ranging from scientific/medical visualization to information visualization and dimensionality reduction. Our key contributions are:

- Introduce AVAs, a new paradigm that leverages the visual perception capability of a machine learning model for autonomous decision-making in visualization. Make the first step toward building visualization agents that can act as virtual visualization experts;
- Provide a preliminary exploration of state-of-the-art multimodal LLM’s visual perception ability for interpreting different visualization outputs, including, scatterplots, parallel coordinate plots, graphs, volume rendering outputs, etc.
- Demonstrate the feasibility and wide adaptability of AVA on several distinct visualization applications.

2. Related Works

2.1. Visualization Generation and Recommendation

Several existing tools explore how to generate visualizations based on user instructions. Data2vis [DD19] utilizes a recurrent network to generate code for visualization (e.g., with Matplotlib [WZW*23], or VTK [SK19]). The NL4DV [NSS20] approach turns visualization queries into visualization descriptors within the Vega-lite grammar [SMWH17], and the work by Mitra et al. [MNES22], explores the back and forth interaction with such visualizations using a natural language interface. LLMs have been adopted for a similar role in LIDA [Dib23]. The KG4Vis work [LWZ*21] adopts knowledge graphs to produce visualization recommendations. Besides just generating the visualization, several methods explore utilizing machine learning (ML) to help explain the rationale behind why a given visualization is recommended. For example, AdaVis [ZWLQ23] adopts an attention-based model for explainable visualization recommendation, whereas the follow-up work leverages LLMs [WZW*23] to achieve a similar goal. With all such adaptations of LLMs in the visualization pipeline, it is also important to understand the width and depth of their capabilities beyond more constrained problems and contexts. Chen et al. [CZW*23] evaluate LLMs for solving visualization coursework by directly feeding them assignment descriptions. Apart from generating code that produces visual output, LLMs are ideal for text description generation. Zong et al. [ZLL*22] utilize LLMs to generate descriptions of visualization for visually impaired users to understand and navigate the visualization.

Compared to these previous works, we are utilizing LLM in the proposed work in two significantly different ways. Firstly, our primary goal is to develop an agent that can refine visualization iteratively to accomplish specific visualization tasks, the objective of the agent is to understand and build upon an existing visualization, rather than focus on generating the visualization in the first place. In other words, we make the role of the agent as a visualization user rather than a visualization designer. Second, all existing works do not utilize the visualization output as input for the ML system for subsequent analysis, which significantly limits the capability and flexibility of the system. To the best of our knowledge, the proposed AVA is the first work that utilizes the visual perception of a multimodal LLM for visual analysis and autonomous decision-making.

2.2. LLM-based Autonomous Agents

With recent advances in LLM, there has been an explosive interest in developing LLM-based autonomous agents. Compared to traditional reinforcement learning agents that often need to develop world understanding from scratch, LLM’s in-depth prior knowledge and information processing capability make them more adaptive to complex environments and solving intricate tasks. Voyager [WXJ*23] introduces an LLM-powered embodied agent in Minecraft that can continuously explore the world and achieve milestones in the game world that were not possible with previous reinforcement learning approaches. Due to the vast literature in this space and relevance to the current work, we refer readers to a comprehensive survey on LLM-based agents [WMF*23]. In the

following discussion, we will focus on vision task agents-related works.

Even though most LLM models are not designed from the ground up for processing visual inputs, many recent works try to incorporate external vision model [SMV23] or develop auxiliary components and fine-tune the model to provide additional vision capability [FZF*23, ZSC*23]. ViperGPT [SMV23] developed an agent that is capable of dividing the task into individual API calls to an external vision model for answering image-based queries and beyond. The layoutGPT [FZF*23] work introduced visual understanding for generating and reasoning about object placement in images and 3D scenes. Gpt4roi [ZSC*23] augment LLM for fine-grained spatial reasoning capabilities.

These adaptations often focus on specific tasks and are trained on smaller-scale data, therefore are not designed for more general capabilities. This changes with the recent introduction of the GPT4-V (vision) [Ope23] model by OpenAI, which added visual perception to one of the largest and most capable LLM models. A detailed evaluation of a broad spectrum of visual understanding tasks is discussed in the “The Dawn of LMM” work [YLL*23]. Compared to the more general evaluation task in [YLL*23], we try to explore the GPT4-V model’s visual perception capabilities on a specially designed set of visualization tasks, and eventually design an agent that is capable of refining and improving visualization output autonomously.

3. Background on Multi-modal Models and LLM Agents

Multi-Modal LLM. Recently, we saw increasing popularity for models that can take multiple modalities as input [RKH*21] or models with input in one modality while output in another modality (e.g., text-to-image [RBL*22, RDN*22], and image-to-text [YVW*22]). In the context of this work, we focus on multi-modal LLMs [YFZ*23] that can take both image and text as input. LLMs, due to their capability and scale both in terms of parameters and training data size, are often referred to as foundation models. These foundation models have access to a broad range of knowledge for understanding implicit context and common sense that was not possible before. Since humans interact with the environment through multi-modal sensory input, the multi-modal LLM is an inevitable evolution of the text-only LLM systems. At the time of writing, the state-of-the-art multi-modal LLM is the GPT4-V (Vision) model, which the proposed work has utilized. However, it is important to note that the proposed AVA is not tied to a specified implementation of multi-modal LLM.

LLM Agents. In the context of machine learning, we can refer to an agent as a system or program that can autonomously make decisions or perform actions based on its environment. An agent should be able to take action and make observations of its environment, and then reason about these observations before deciding on the subsequent action. The action of the LLM agent can be in the form of textual or image output, however, their capabilities can be greatly enhanced by allowing them (e.g., ToolLlama [QLY*23]) to make direct API function calls to utilize external tools. When utilizing an LLM as the brain of the agent, we can enhance its action planning capabilities by adopting some simple protocols, such as

chain-of-thoughts [WWS*22] (i.e., instruct it to solve the problem “step-by-step”), or ReACT [YZY*22] that connection the reasoning process with action the agent can take.

4. Preliminary Exploration of Multimodal-LLM for Static Visualization Perception

Before we can design an effective visualization agent that relies on visual input for decision-making, it is crucial to obtain some basic understanding regarding its capabilities and limitations for perceiving various types of visualization output. It is important to note that this is **not intended to be a rigorous and systematic evaluation of multi-modal LLM ability**, as an in-depth study will require substantial resources and effort that is beyond the scope of this work. We hope this assessment can help illustrate what type of visualization agents we can realistically design and what would be the ideal tasks for such agents.

Volume Rendering. We begin with evaluating the LMM’s ability to recognize structures of interest within direct volume rendering images. Unlike photo-realistic images, that was explored by the work of Yang et al. [YLL*23], the outputs of volume rendering are subject to additional complexity (i.e., varying transparencies) introduced by the underlying transfer function. To assess the model’s capability, we present the model with the task of examining a screenshot and determining whether a specific object or structure of interest is ‘recognizable’ or ‘not recognizable’. We define the two assessments for the prompt as follows: **recognizable**: *The structure of interest and its shape can be discerned in the screenshot.* **not recognizable**: *The structure of interest cannot be identified in the image, even if another structure is recognizable.*



Figure 3: The Boston Teapot dataset volume rendered using the same color map but at varying opacity levels. Structure of interest: the teapot. The response from the LLM model was 18a: ‘not recognizable’, 18b: ‘recognizable’, 18c: ‘recognizable’, and 18d: ‘not recognizable’

We assess the model on two datasets, the Boston Teapot [TIH], and a downsampled version of the Visible Male [SASW96]. The reason why the Boston Teapot was selected for this experiment is because there is another structure, a lobster, located inside the teapot. As illustrated in Figures 18, we maintained fixed rendering parameters, including viewpoint and colormap, while introducing variations solely in the opacity transfer function. To maintain uniformity across experiments, we employed a fixed-width (1/10th of the value range) triangular function for the opacity transfer function, altering only the peak value in the center of the window. As shown in the Figure, the model consistently provided accurate assessments in all cases. More details on the experiment and other assessments for other datasets can be found in the supplementary material.

Scatterplot. Compared to volume rendering images, in which visual recognition is a simple binary task, i.e., object recognition, the assessment of visual structure in scatter plots is more nuanced. Here, we design five basic visualization tasks: *cluster recognition*, *cluster counting*, *outlier detection*, *outlier counting*, and *correlation detection* to evaluate its performance. The evaluation result is displayed in Table 1. For cluster recognition, our experiments show that the model can easily tell the plot has clusters (100% success rate). However, for the counting task, the success rate of the model is only at 60%. Similarly, we also separate the outlier tasks into recognition and counting. The final result aligns with the cluster recognition task. The model performs well on the outlier recognition task which has a 100% success rate, but has medium performance on the outlier counting task. In the correlation detection task, the model has a 100% success rate.

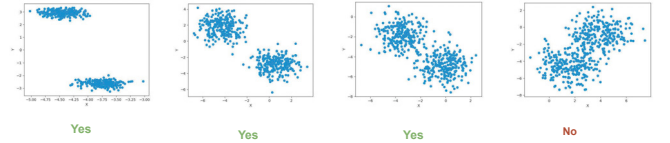


Figure 4: The ability of GPT4-V to identify clusters in the scatter plot with different levels of ambiguity.

The experiment results show that the model has a decent ability to understand and analyze the scatter plot. However, in the experiment, most of the visualizations have clear signals to tell whether certain features exist. This raises another question of whether the model is able to identify ambiguous cases. We perform another simple experiment in which a scatter plot has two clusters but with different point spreads. From the result of Fig 4, we can tell that except for the last one which it is hard for humans to tell whether it has two clusters or not, the LLM model is able to identify the rest of the example accurately.

Tasks	scatter plot(success rate)	parallel coordinates
cluster	100%	100%
cluster count	60%	20%
outlier	100%	90%
outlier count	60%	80%
correlation	100%	20%

Table 1: The performance of GPT4-V on a scatter plot and parallel coordinate tasks. GPT4-V can identify outlier and cluster well in both visualizations. However, its ability for object counting is comparatively poor. Meanwhile, the correlation detection in parallel coordinates plot is also limited.

Parallel Coordinates Plot. We examine parallel coordinate plots with the same tasks as the scatter plot. Both experiments have a similar setup on cluster and outlier tasks, except the number of dimensions in each dataset will change from 2D to 5D. The overall results are a bit worse than the model’s performance on scatterplot visualization. In cluster and outlier recognition tasks, the model performs well. In the cluster counting task, parallel coordinates perform badly with a 20% success rate but in the outlier counting task,

the GPT4-V model performs well. Opposite to the correlation task, the parallel coordinate makes it hard to identify correlation relationships.

Tasks	node count	find node	connection	neighbor
success %	50%	100%	70%	10%

Table 2: The performance of GPT4-V on common graph tasks.

Graph. To assess GPT4-V’s visual understanding of graphs, we choose the classic graph visualization technique node-link diagram and adjacency matrix. In our experiment, we use the basic graph exploration task [GFC04] to evaluate the performance of the LLM. Instead of performing all tasks, we pick four tasks that are easy to perform without interactions. The overall result is displayed in Table 2. From the evaluation, we can tell that LLM can easily find a node in the graph visualization. However, it is difficult to tell the neighbor of the selected node. The connection tells whether two nodes are connected (directly or indirectly through other nodes), and the final result shows sub-optimal performance. Finally, the node count ability has a 50% success rate which shows that the model again has poor performance on the counting tasks.

Despite the relatively limited exploration, our experiment demonstrates the model’s capability to discern structures and objects in volume rendering results. Among the information tasks, the model achieves better performance on scatterplots compared to parallel coordinate plots or graphs. Therefore, to leverage the strength of the system, in our case study (see section 6), we focus on volume rendering and scatterplot-related applications.

5. Autonomous Visualization Agent (AVA)

We define AVA as a paradigm for designing AI-driven agents that serve as a medium between a specialized visualization tool and a domain user. The key principle of AVA involves the utilization of machine vision for decision-making. It takes user instruction in natural language and achieves the user-specified goal by operating the visualization tool autonomously based on the visual understanding of visualization outputs. And we refer to the concrete implementation of AVA as AVAs.

5.1. Key Components of AVAs

To achieve its design goal, the AVAs need to accurately perceive visual input and make plans on what action to take based on current visualization results and do so by following user natural language instructions. As illustrated in Figure 5, AVAs need to contain at least three key components, namely visual perception, action planning, and memory.

Visual Perception the visual perception is at the center of the AVAs’ capability, and what distinguishes it from existing LLM applications in visualization. There is some similarity between AVA and an embodied agent [ZDS*23] in robotic research, where an agent will take action based on sensory input (e.g., vision) and observe the impact of the action in the environment. Similarly, for AVAs the sensory input is the visualized image, and the action corresponds to changes in the visualization setting (e.g., update parameters), and the impact of the action is a new visualization output.

Action Planning In order to make autonomous decisions and respond to the “sensory” input, the AVAs need an action planning component as the “brain” of the system. Here we have a range of choices for its design. As illustrated in Figure 2, we can either rely more on heuristics to drive the action planning or let the LLM do everything on its own, which corresponds to two distinct approaches to the action planning design.

- **Heuristic-Centric:** infuse our existing domain knowledge into heuristics for how to update the visualization tool based on assessment from the visual perception component. Their action plan is defined explicitly. In such a scenario, the visual perception and assessment essentially act as a loss function for a pre-defined optimization procedure.
- **LLM-Centric:** leverage the capability and prior knowledge of LLM to guide the exploration of the action space. Their action is only influenced by the initial prompt feed to the system.

One important thing to note is that both approaches will provide autonomous decision-making based on visual perception, so from the user’s perspective there may be little difference. The distinct between them comes from whether we want to rely on our own prior knowledge explicitly or we hope to leverage the LLM’s capability for planning and suggestion, while only influencing its behavior indirectly.

Memory Beside the visual perception, and action planning components, the other essential part of AVA that both of these components need is memory of the previous actions or the visualization outputs it observed before. In order to make complex and well-informed decisions, we often need to refer back to or compare with previously examined results or conclusions. The same is true for AVAs.

Visualization-Perception-Action Loop Besides the three key components, one essential aspect of AVA, and its key capability, is associated with the autonomous visualization loop, i.e., from visualization to perception and then to action. The process is bootstrapped by the specific high-level task given by the user and starts with a default visualization setup, and then the system:

- Generating visualization output by executing API calls to the visualization tool based on the given parameter.
- Leveraging the visual perception component to comprehend semantics and structure in the current visualization.
- Provide assessments of whether the visualization achieves the user-set goal, and the action planning component makes decisions on what visualization parameters it should use next.

The agent will iterate through these steps until the visualization goal is achieved. This methodological framework forms the foundation of AVAs, enabling us to utilize the visual perception system and the optimization strategy of LLMs or heuristics to interact with visualization outputs effectively. It addresses the critical need for autonomous agents capable of navigating complex visualization tasks with precision and adaptability.

5.2. Implementation

So far, we have discussed the conceptual idea of how AVA works. In this section, we provide practical guidance on their implementation. Our implementation utilizes the GPT-4 Vision model [Ope23]

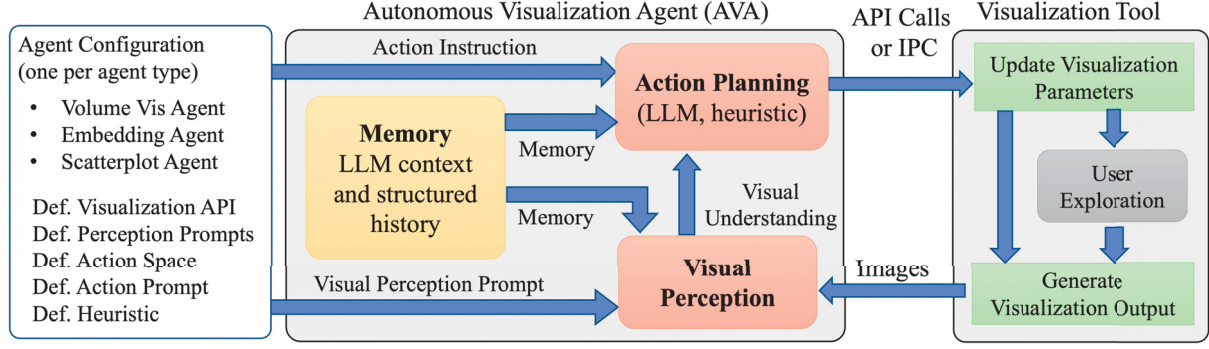


Figure 5: The breakdown of the components of the AVA. The capability of the AVAs hinged on their visual perception, and the visual understanding can then be utilized by the action planning system to modify/steer the visualization tool. In order for AVAs to make informed decisions and the ability to understand context, they also need a memory component that both visual perception and action planning components can easily access.

for visual perception and action planning (for LLM-centric scenario), harnessing its natural language understanding capabilities alongside a visual perception engine.

To establish a flexible and reusable foundation for our AVAs, we have created an abstract class that acts as the basis for any specific type of agent. It contains several core functionalities: 1) a unified interface for accessing LLM API for visual perception or action planning tasks; 2) a basic blueprint on how an agent should interface with the visualization tool; 3) configuration functionality that helps define the agent, e.g., prompts template; 4) capability to parse and extract visual assessment results, parameter, and function call information from the language response.

concrete classes that inherit the base one, and in the new class, application-specific logic, e.g., heuristics-centric action planning, can be implemented. Each of the concrete classes will also have an associated JSON configuration file, prompts or part of prompts are organized in a structured fashion. Our AVA implementation is in Python. As illustrated in Figure 6, we design a simple layout for the interface, with the control and conversion history on the left panel and the visualization of interests on the right.

Agent Initialization To initiate an AVA’s functionality, we establish a context by prompting the Large Language Model (LLM) with the assumed role of the agent. This definition typically encompasses several elements: scenario, visualization task, goal, approach, and constraints. The prompt structure typically follows this format:

You are an autonomous visualization agent tasked with assisting a user in {visualization task}. In each step, you will receive a screenshot and you will assess the image and provide the {approach}. Your goal is to determine {goal}. Achieve this goal by {approach}, adhering to the following constraints: {constraint 1, constraint 2...}

One of the most critical components of AVA implementation is creating natural language prompts. Crafting effective prompts is essential for defining the agent’s role and specifying the approach to achieving the visualization task.

Connection Between AVA and Visualization Tool The AVA can either directly call the API, provided both the agent and the visualization run in the same application/context. However, to maximize the flexibility and support complex external tools (e.g., GPU accelerated direct volume rendering), we also include support for a more generic solution with inter-process communication (IPC) mechanisms to facilitate seamless data exchange between the agent and the visualization tool. In our implementation, we utilize the RPyC (Remote Python Call) [Fil13] to facilitate the IPC.



Figure 6: AVA integrated visualization interface. We implement AVA as a configurable system where individual components can be combined/configured for different target applications.

For specific types of applications, AVAs can be developed as

Agent-driven Assessment and Opacity Transfer Function Iterations

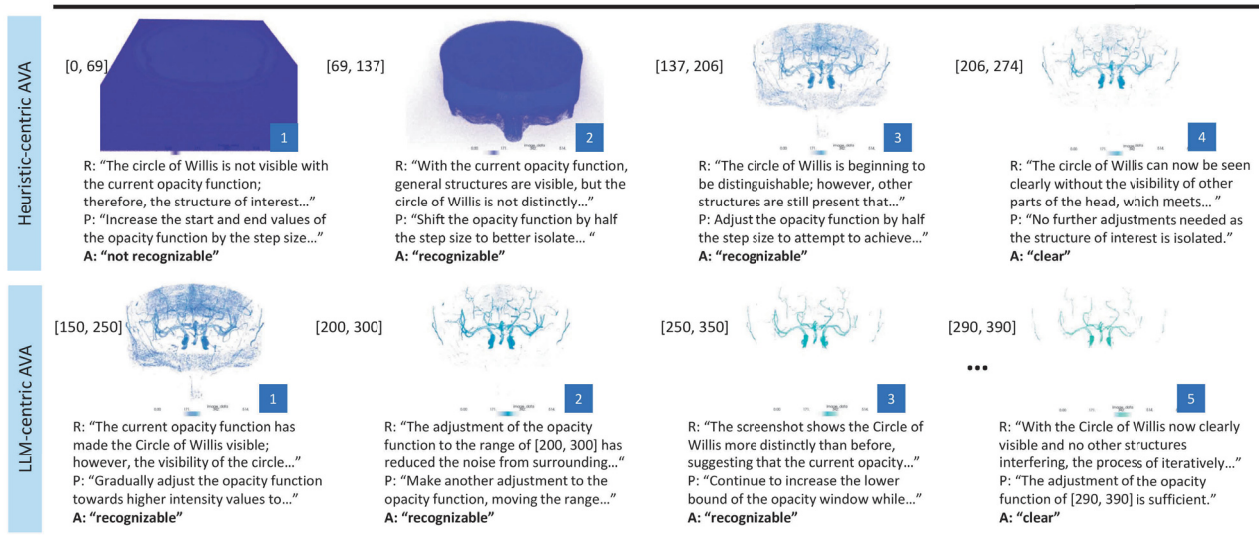


Figure 7: The results from the heuristic-centric and LLM-centric AVAs. The screenshots are generated from the proposed opacity transfer function. The response includes (R)easoning, (P)lan, and (A)ssessment by the agent, and based on this, the agent suggests a new pair of values to construct the triangle-shaped opacity transfer function. Each agent converged towards an opacity function rendering the structure of interest.

6. Case Studies

6.1. Opacity Transfer Function for Volume Rendering

In this case study, we focus on the opacity transfer function design process—a crucial task in volume rendering, where structures of interest must be appropriately depicted within the opaque range of the opacity transfer function. We test the agents with a dataset of a head [hea], which is a 3T Time-of-flight Magnetic Resonance Angiography and contains part of the portion around the height of the eyes where the brain arteries are located. It contains skin, soft tissues, the skull, and the vascular structure inside. The most interesting structure inside this dataset is the arterial blood supply of the brain, called the circle of willis.

Scenario: Agent assists the user in volume rendering.

Visualization Task: Evaluate the visualization output and determine the appropriate opacity transfer function for rendering a structure of interest within a volume.

Goal: Identify an opacity transfer function that accurately renders the structure of interest.

To facilitate a comprehensive discussion of AVA behavior, we implemented two different agents. A heuristic-centric agent receives the action plan as a heuristic defined by the user, while the LLM-centric agent utilizes the model’s knowledge about the opacity transfer function design in order to devise a strategy. For both agents, the opacity function remained a triangle function with the peak value positioned between the start and endpoints. The viewpoint and color map were also fixed. The AVAs provide assessments categorized as ‘recognizable’, ‘not recognizable’ as described in Section 2. In addition, we added the ‘clear’ assessment, as a stopping criterion for the optimization, which denotes that the

structure of interest is distinctly visible without any other structures occluding it. This assessment is necessary for the AVA in order to improve results upon finding an opacity value range that renders the structure of interest ‘recognizable’, which could still contain a large amount of noise. The agents are described as follows (details are available in the supplementary material).

Heuristic-Centric: In this setup, the agent provides assessments, but the opacity transfer function adjustments are defined by the agent designer. For a proof-of-concept, we utilize a simple linear search-based approach that shifts the window of the opaque range towards higher values, while the function always assumes a triangle shape. For these tests, we selected parameters to separate the value range into 10 bins, where the window width is one bin wide. We shift the window one bin with each iteration. We also added a fine-tune parameter, where we reduce the speed when the structure of interest is ‘recognizable’ but it is not yet ‘clear’. In that case, the window shifts only by half of its width. The only information the domain user needs to provide here is the structure of interest (the circle of Willis, a vascular structure in the brain) and the value range.

In general, this action plan can be implemented in two ways, depending on the scenario. This heuristic can be added as code, such as a plugin integrated into the visualization code. To elucidate the agent’s behavior in alignment with the heuristic described above, the agent initially employs a triangle function at the far left end of the value range and incrementally moves upward. It continues this process until it can confidently recognize the structure of interest, in this case, the circle of Willis. As the agent recognizes the vascular structure, it takes half a step to make smaller adjustments to the opacity transfer function until it can fully discern the circle of

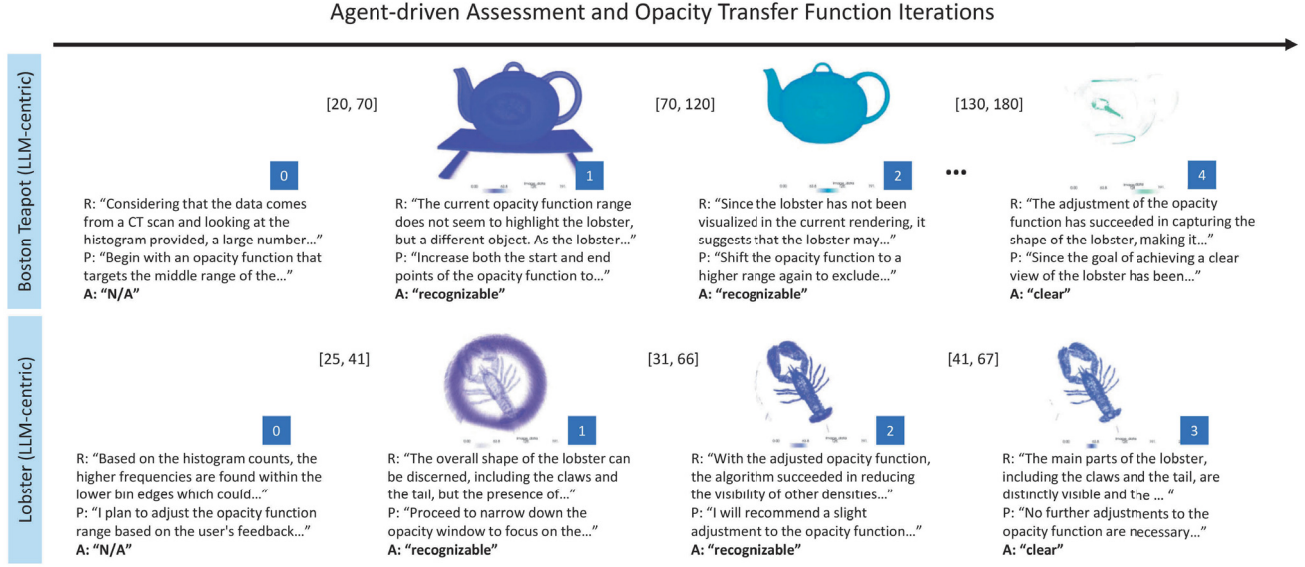


Figure 8: The results from the LLM-centric AVA for the Boston Teapot and the Lobster in Resin datasets. The structure of interest is *the lobster* in both cases. The rendered images are generated from the proposed opacity transfer function. The response includes (R)reasoning, (P)lan, and (A)ssessment by the agent. In the top row, the agent suggested the first opacity transfer function that revealed the teapot instead of a lobster and it moved on higher value ranges and successfully detected the lobster, even at a low resolution. In the bottom row, AVA found the lobster, which has higher resolution and occupies a larger space in the volume, and immediately and fine-tuned the result until almost no resin was visible in the visualization output.

willis. While we used a linear search strategy in this demonstration, other approaches, such as binary search, can also be employed. An advantage of utilizing large language models (LLMs) is the ability to incorporate text from research papers on opacity transfer function design, providing the agent with a more advanced action plan.

LLM-Centric: This AVA is not limited to a user-defined heuristic for adjusting the opacity function. Instead, it can leverage the prior knowledge of transfer function design inherent in the LLM to facilitate the design process. However, it remains constrained to providing a triangle function as the opacity transfer function. We provided the agent with the acquisition modality and the histogram to provide it with similar information as a human user would have. As depicted in Figure 7, the agent explores various opacity ranges until it successfully generates the function for the circle of willis. Notably, in this case, the agent operates with a greater degree of autonomy, employs a strategic approach, and reflects on past decisions as explained in its "reasoning" and "plan" as shown in Figure 7. Interestingly, it immediately devised a plan, where it starts with a range higher than the first peak in the histogram, which it correctly assumes is the background.

We further tested the AVA’s capabilities on structures that are more challenging to find an appropriate opacity function. Specifically, we utilized the Boston Teapot dataset discussed in Section 4, which contains a lobster inside the teapot. The lobster in this dataset can be only partially visualized due to the low resolution of the data. This presented a more difficult scenario compared to the circle of willis. Additionally, the lobster is relatively small within the dataset, resulting in its representation by a very low bin in the

histogram. Despite these challenges, as demonstrated in Figure 8, the agent successfully determined the correct opacity transfer function within a few steps when tasked with identifying the lobster structure of interest. Following the reasoning in each step, it reveals its advanced action-planning capabilities. To provide a comparison, we also tested the agent on another dataset, containing a Lobster in Resin ($301 \times 324 \times 56$, *uint8*, Courtesy of VolVis distribution of SUNY Stony Brook, NY, USA.). In both cases, the model reasoned that the second histogram peak might be the structure of interest, however in the Boston Teapot dataset, it found the teapot instead of the lobster. Remarkably, the LLM-centric agent moved on and tried different value ranges and found the lobster in just a few steps, even though the lobster is harder to recognize due to the low resolution. In comparison, in the Lobster in Resin dataset, the lobster was revealed together with the resin in the first iteration and then the agent fine-tuned the opacity in Step 3 until no resin is visible anymore.

This successful demonstration illustrates the agent’s robustness in handling challenging scenarios and its ability to swiftly adapt to different datasets and structures of interest. The detailed results of these agents and their responses are provided in the supplementary material.

6.2. Scatterplot Opacity Optimization

Apart from rendering output, from our initial assessment (section 4) the GPT-V has better visual perception for scatterplot compared to all other common information visualization encodings (e.g., parallel coordinate, graph). Therefore, we focus the rest of the case

studies on the scatterplot type of visual output. In this section, we examine the optimization of opacity value for scatterplot points to mitigate the occlusion effects from overplotting.

The perceptual base opacity optimization has been explored in the visualization domain [MAF15, MPOW17], either through a data-driven modeling perspective based on user preferences [MAF15] or through a visual perception modeling approach by designing a cost function that captures relevant aspects of the human visual response [MPOW17]. Here we do not aim to directly compare with these existing methods, as a meaningful comparison requires an extensive and controlled study. We hope to use this case study to illustrate how a fundamentally different approach to address the opacity optimization challenge can be obtained by a straightforward adoption of the AVA framework. From the existing study on user preference [MAF15], the relationship between the point opacity and assessment of overplotting level follows an inverse logarithmic relationship, i.e., overplotting only gets better when point opacity gets much lower in the 0.0, 1.0 range. This is crucial prior knowledge that should be incorporated into the design of the agent. Therefore, we adopted the heuristic-centric approach outlined in Section 5, where we encode the logarithmic relationship into our search procedure. At the start of the optimization, we set the initial opacity $O = 1.0$. The floor opacity, i.e., lowest allowable opacity $O_f = 0.0$. For each step, we will update the new opacity as $O' = O_f + (O - O_f)/2$, essentially half the opacity value different between the current opacity and the floor opacity. By providing the model with scatterplot images generated with opacity O' and O , we then evaluate which opacity is better suited for the given data. If the new opacity is deemed too low, we then set it as the new floor opacity O_f . We continue to iterate to narrow down the selection until the opacity different threshold is reached.

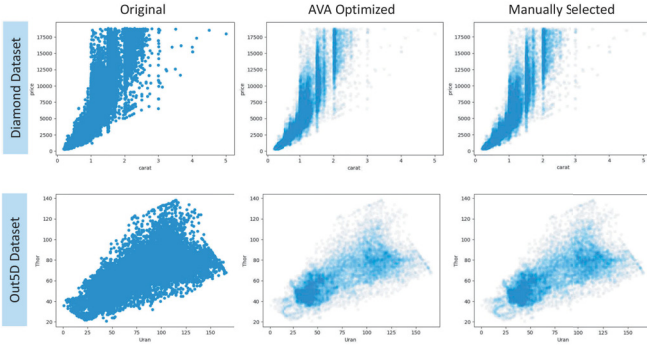


Figure 9: Scatterplot opacity optimization results. The left column shows the original plots with severe overplotting, the middle column shows the agent optimization results, and the right column shows independent manual selection results, there are some minor differences, but the overall results are comparable.

As shown in Figure 9, each row indicates a different dataset, namely Diamond data [kag], and the Out5D data [xmd]. The first column includes the original scatterplots with the overplotting issue when opacity is 1.0, the middle row is the AVA opacity-optimized scatterplots, and the last row is the human user reference obtained independently from the optimization interface. As we can see, the AVA-generated scatterplot closely matched the user preference.

6.3. Dimension Reduction Hyperparameter Tuning

The choice of hyperparameters can greatly impact t-SNE [VdMH08] and UMAP [MHSG18] results. Inappropriate hyperparameters may lead to misleading interpretations of the high-dimensional structure, and they often need to be tuned for a given dataset. Here, we utilize AVA to perform automatic hyperparameter tuning for identifying more suitable hyperparameters, for both single-hyperparameter and multi-hyperparameter cases. Considering the prior knowledge the LLM is likely to have on these common methods, we opt for LLM-centric action planning, where the LLM directly suggests hyperparameters. In Figure 10, we show the single parameter optimization result, where we only optimize the most sensitive parameter for each method, i.e., perplexity for t-SNE, and the neighborhood size for UMAP. We withheld the class label from the agent to use as the ground truth for evaluation. All plots are generated from the data RNA sequence data [TYG*18] with 20 classes. As we can see for the UMAP embedding, the default parameter gives a small number of stringy clusters (a), whereas, in the optimized embedding (b), several classes that were linked together are now separated. For the t-SNE case, there is a less clear advantage for the optimized embedding in terms of cluster separability, however, it does show more compact clusters.

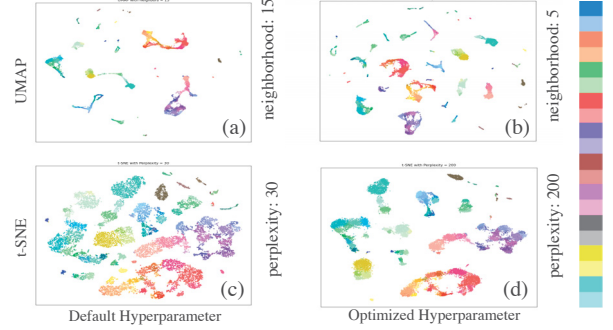


Figure 10: Hyperparameter optimization results. For UMAP, the optimized hyperparameter (b) shows greater class separation compared to the default (a). For t-SNE, (d) shows a similar class separation but with more compact clusters.

Our experiment with the multi-hyperparameter (up to 5) agent case, however, is largely unsuccessful. As we see the suggested hyperparameters bounce back and forth during the optimization process. This indicates the potential challenge for the agent to explore higher dimensional action space (see more detail in Section 8).

7. Experts Feedback on AVA

Given that we did not aim to conduct a general evaluation of multi-modal foundation models, nor did we target a specific application area, we chose to gather informal feedback from experts in key domains that could provide valuable insights into the potential of these agents for visualization tasks. Our feedback collection process involved two senior AI researchers, a professor in medical visualization, and a professor who heads an Institute for Radiological Diagnostics and Intervention. The latter two were selected due to their daily workflow experience with volumetric data and general

visualization tasks. For all the feedback sessions, we first demonstrated the use cases of AVAs as described in Section 6, and then conducted unstructured interviews with the experts. While the discussions were mostly open-ended, we did inquire about their overall feedback on how such visualization agents might impact their workflow, and where the potential benefits and limitations of this paradigm are.

The medical visualization expert said: *“Their ability to comprehend visual elements and identify structures is indeed impressive, laying a solid foundation for the future development of such agents. With this substantial potential at hand, the utilization of these agents now lies in the hands of visualization researchers, who have the opportunity to harness their capabilities for innovative applications.”* She extended the discussion by suggesting the creation of a generic workflow that can incorporate how visualization experts use the volume rendering tool into the agent prompts to enhance their capabilities.

The head of the radiology institute said: *“I’m impressed with the semantic understanding, reasoning capabilities, and high autonomy exhibited by the agents. There is exciting potential to replace trivial visualization tasks that until today require a radiologist.”* He envisioned the use of these agents for double reading in radiology, where two independent radiology reports could be generated to cross-validate diagnoses. However, he remained skeptical about whether the recognition could go beyond simple shapes, for example actually perceiving differences between arteries and veins, as well as extending the visual perception capability to make assessments based on multiple image modalities.

The senior AI researchers said: *“1: This can be a very general approach. One additional application I can see this working is for finding more informative views for 3D plots, which I always have trouble with.” “2: The AVA setup can be easily extended to other types of user interfaces beyond just visualization. One thing I am interested to know is how well it handles a larger action space, will the search fail or converge?”* The AI experts believe this is a fundamentally different way to think about data visualization problems and see the connection with embodied agent research. One potential concern they mentioned is whether the action planning can work with a much bigger action space. As a response to their feedback, we extend our dimensionality reduction case studies to include additional experiments with up to 5D space. Overall, the feedback from all the experts underscores the transformative potential of AVA.

8. Discussion and Future Work

In this work, we investigated the capabilities of emerging multi-modal foundation models like GPT-4V and their ability for visualization tasks. Building on these findings, we introduced a template for a novel paradigm known as Autonomous Visualization Agent (AVA) for solving high-level visualization tasks through visual perception and action planning. Despite demonstrating its feasibility through our case studies, it is also essential to acknowledge their limitations.

Prompts Engineering. With the flexibility and usability of natural language, it also brings certain limitations. The precision of AVAs

heavily relies on the choice of the prompt, as the natural language remains inherently fuzzy and context-dependent. This limitation may diminish in significance as these models evolve to become more powerful and context-aware. This challenge can also be partially mitigated through heuristic-centric action planning through explicitly coded search logic, nevertheless, the visual perception still relies on prompts to convey the assessment objectives.

Large Action Space. The action planning component of AVA is essentially doing an exploration of a potentially high-dimensional action space. The search is guided by the visual assessment, which theoretically can be considered as the loss in a zeroth order optimization scenario [SM22]. However, despite having a black box loss, we as visualization designers or the LLM do have prior knowledge of the action space that could guide the exploration to allow a fast convergence. Still, as the problem is associated with the degree of freedom of the underlying system, there is no easy solution, we believe LLM model with stronger prior (i.e., visualization task fine-tuned models) and better optimization strategy is likely required to allow AVAs to reach their full potential.

Future Directions. Our plan for future work involves a more extensive evaluation of the current models’ capabilities in understanding visualization output, expanding on the foundation laid in Section 4. This evaluation will provide deeper insights into the extent to which multi-modal foundation models can contribute to visualization research. Additionally, we plan to explore different agent setups, including increasing the number of agents and increasing the interactivity of the agent. By implementing multiple independent agents with slightly different definitions, we can offer a means of cross-validation for applications with low error tolerances. So far, we have demonstrated agents employing a closed-loop optimization strategy with intermittent communication with the user. By tuning the level of interactivity, as in a chatbot, we could create an even tighter symbiosis between a human expert and an AI for a joint visualization task. For additional application scenarios. There are many possibilities, as mentioned by one expert we interviewed, adjusting the viewpoint to avoid visual occlusion in 3D visualization or 3D plots can be a great use case. They can be particularly useful for offline rendering in HPC applications and large-scale data etc. The model’s ability for scatterplot understanding can also be utilized to design customized diagnostics metrics for exploratory data analysis.

9. Conclusion

The primary objective of this work is to underscore the significance of autonomous visualization agents in enhancing the accessibility of visualization tools. We have demonstrated that not only are these agents possible, but they can already be useful for solving non-trivial visualization tasks. As multi-modal foundation models continue to advance in power and sophistication, we anticipate a corresponding increase in the capabilities of such agents. In many ways, we are speculating on how the ongoing development of Large Language Models (LLMs) can reshape the landscape of visualization research. The fusion of image understanding and language understanding within these multi-modal foundation models holds the potential to fundamentally transform the way we think about visualization and user interaction. With further development, we believe

AVAs can eventually serve as virtual visualization experts in the room, streamlining the entire visualization pipeline beyond the automatic visualization parameter adjustments that have been demonstrated in this paper. In conclusion, our research opens exciting possibilities for the future of visualization tool design that aims at the collaboration between humans and AI-driven agents.

Acknowledgement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. The project is supported by LLNL LDRD (23-ERD-029, 23-SI-003). The work is reviewed and released under LLNL-CONF-857838. We thank the experts Prof. Renata Raidou (TU Wien), Prof. Christian Nasel (Medical University Vienna), Dr. Jayaraman Thiagarajan (LLNL), and Dr. Rushil Anirudh (LLNL) for their valuable feedback on the capabilities of our agents.

References

- [CZW*23] CHEN Z., ZHANG C., WANG Q., TROIDL J., WARCHOL S., BEYER J., GEHLENBORG N., PFISTER H.: Beyond generating code: Evaluating gpt on a data visualization course. *arXiv preprint arXiv:2306.02914* (2023). 3
- [DD19] DIBIA V., DEMIRALP Ç.: Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE computer graphics and applications* 39, 5 (2019), 33–46. 2, 3
- [Dib23] DIBIA V.: LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (Toronto, Canada, July 2023), Association for Computational Linguistics, pp. 113–126. URL: <https://aclanthology.org/2023.acl-demo.11>, doi:10.18653/v1/2023.acl-demo.11. 2, 3
- [FAT99] FUJISHIRO I., AZUMA T., TAKESHIMA Y.: Automating transfer function design for comprehensible volume rendering based on 3d field topology analysis. In *Proceedings Visualization'99 (Cat. No. 99CB37067)* (1999), IEEE, pp. 467–563. 2
- [Fil13] FILIBA T.: Rpyc (remote python call), 2013. Python library for remote procedure calls. URL: <https://rpyc.readthedocs.io/en/latest/>. 6
- [FZF*23] FENG W., ZHU W., FU T.-J., JAMPANI V., AKULA A., HE X., BASU S., WANG X. E., WANG W. Y.: Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393* (2023). 3
- [GFC04] GHONIEM M., FEKETE J.-D., CASTAGLIOLA P.: A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization* (2004), pp. 17–24. doi:10.1109/INFVIS.2004.1. 5
- [hea] Head dataset. <http://www.celebisoftware.com/Dataset.aspx?catId=3>. Accessed: [Your Access Date Here]. 7
- [Hun07] HUNTER J. D.: Matplotlib: A 2d graphics environment. *Computing in science & engineering* 9, 03 (2007), 90–95. 2
- [kag] Kaggle diamond dataset. <https://www.kaggle.com/datasets/shivam2503/diamonds/data>. Accessed: YYYY-MM-DD. 9
- [LKG*16] LJUNG P., KRÜGER J., GROLLER E., HADWIGER M., HANSEN C. D., YNNERMAN A.: State of the art in transfer functions for direct volume rendering. In *Computer graphics forum* (2016), vol. 35, Wiley Online Library, pp. 669–691. 2
- [LWZ*21] LI H., WANG Y., ZHANG S., SONG Y., QU H.: Kg4vis: A knowledge graph-based approach for visualization recommendation. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 195–205. 3
- [MAF15] MATEJKA J., ANDERSON F., FITZMAURICE G.: Dynamic opacity optimization for scatter plots. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), pp. 2707–2710. 9
- [MHS18] MCINNES L., HEALY J., SAUL N., GROSSBERGER L.: Umap: Uniform manifold approximation and projection. *Journal of Open Source Software* 3, 29 (2018), 861. 9
- [MNE22] MITRA R., NARECHANIA A., ENDERT A., STASKO J.: Facilitating conversational interaction in natural language interfaces for visualization. In *2022 IEEE Visualization and Visual Analytics (VIS)* (2022), IEEE, pp. 6–10. 3
- [MPO17] MICALLEF L., PALMAS G., OULASVIRTA A., WEINKAUF T.: Towards perceptual optimization of the visual design of scatterplots. *IEEE transactions on visualization and computer graphics* 23, 6 (2017), 1588–1599. 9
- [NSS20] NARECHANIA A., SRINIVASAN A., STASKO J.: NL4DV: A Toolkit for generating Analytic Specifications for Data Visualization from Natural Language queries. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2020). doi:10.1109/TVCG.2020.3030378. 3
- [Ope23] OPENAI: Gpt-4 vision. <https://openai.com/research/gpt-4v-system-card>, 2023. Accessed: [Insert date of access here]. 3, 5
- [QLY*23] QIN Y., LIANG S., YE Y., ZHU K., YAN L., LU Y., LIN Y., CONG X., TANG X., QIAN B., ET AL.: Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789* (2023). 3
- [RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695. 3
- [RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3. 3
- [RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763. 3
- [SASW96] SPITZER V., ACKERMAN M. J., SCHERZINGER A. L., WHITLOCK D.: The visible human male: a technical report. *Journal of the American Medical Informatics Association* 3, 2 (1996), 118–130. 4
- [SK19] SULLIVAN C., KASZYNSKI A.: Pyvista: 3d plotting and mesh analysis through a streamlined interface for the visualization toolkit (vtk). *Journal of Open Source Software* 4, 37 (2019), 1450. 2, 3
- [SM22] SLAVIN I., MCKENZIE D.: Adapting zeroth order algorithms for comparison-based optimization. *arXiv preprint arXiv:2210.05824* (2022). 10
- [SMV23] SURÍS D., MENON S., VONDRICK C.: Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128* (2023). 3
- [SMWH17] SATYANARAYAN A., MORITZ D., WONGSUPHASAWAT K., HEER J.: Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)* (2017). URL: <http://idl.cs.washington.edu/papers/vega-lite>, doi:10.1109/tvcg.2016.2599030. 2, 3
- [SWW*23] SONG C. H., WU J., WASHINGTON C., SADLER B. M., CHAO W.-L., SU Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of*

- the IEEE/CVF International Conference on Computer Vision (2023), pp. 2998–3009. 1
- [TIH] TERARECON INC MERL B., HOSPITAL W.: Teapot dataset. <http://www.gris.uni-tuebingen.de/areas/scivis/volren/datasets/data/BostonTeapot.raw.gz>. Accessed: [Your Access Date Here]. 4
- [TYG*18] TASIC B., YAO Z., GRAYBUCK L. T., SMITH K. A., NGUYEN T. N., BERTAGNOLLI D., GOLDY J., GARREN E., ECONOMO M. N., VISWANATHAN S., ET AL.: Shared and distinct transcriptomic cell types across neocortical areas. *Nature* 563, 7729 (2018), 72–78. 9
- [VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-sne. *Journal of machine learning research* 9, 11 (2008). 9
- [WBZ*23] WU Q., BANSAL G., ZHANG J., WU Y., ZHANG S., ZHU E., LI B., JIANG L., ZHANG X., WANG C.: Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155* (2023). 1
- [WMF*23] WANG L., MA C., FENG X., ZHANG Z., YANG H., ZHANG J., CHEN Z., TANG J., CHEN X., LIN Y., ET AL.: A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023). 1, 3
- [WWS*22] WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D., ET AL.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837. 4
- [WXJ*23] WANG G., XIE Y., JIANG Y., MANDLEKAR A., XIAO C., ZHU Y., FAN L., ANANDKUMAR A.: Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291* (2023). 1, 3
- [WZW*23] WANG L., ZHANG S., WANG Y., LIM E.-P., WANG Y.: Llm4vis: Explainable visualization recommendation using chatgpt. *arXiv preprint arXiv:2310.07652* (2023). 3
- [xmd] Xmdvtool homepage. <http://davis.wpi.edu/xmdv/datasets.html>. Accessed: [Access Date]. 9
- [YFZ*23] YIN S., FU C., ZHAO S., LI K., SUN X., XU T., CHEN E.: A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023). 3
- [YLL*23] YANG Z., LI L., LIN K., WANG J., LIN C.-C., LIU Z., WANG L.: The dawn of llms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9 (2023). 3, 4
- [YWV*22] YU J., WANG Z., VASUDEVAN V., YEUNG L., SEYEDHOSSEINI M., WU Y.: Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022). 3
- [YZY*22] YAO S., ZHAO J., YU D., DU N., SHAFRAN I., NARASIMHAN K., CAO Y.: React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022). 4
- [ZDS*23] ZHANG H., DU W., SHAN J., ZHOU Q., DU Y., TENENBAUM J. B., SHU T., GAN C.: Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485* (2023). 5
- [ZLL*22] ZONG J., LEE C., LUNDGARD A., JANG J., HAJAS D., SATYANARAYAN A.: Rich screen reader experiences for accessible data visualization. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 15–27. 3
- [ZSC*23] ZHANG S., SUN P., CHEN S., XIAO M., SHAO W., ZHANG W., CHEN K., LUO P.: Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601* (2023). 3
- [ZWLQ23] ZHANG S., WANG Y., LI H., QU H.: : Adaptive and explainable visualization recommendation for tabular data. *IEEE Transactions on Visualization and Computer Graphics* (2023). 3

Appendix A: Static Visualization Assessment Experiment Setup

Scatter plot

In scatter plot experiments, each task will be performed in 10 experiments, and the final results are aggregated as a percentage number (success rate). In each experiment, 500 points are generated randomly following a pre-defined pattern (e.g., number of clusters).

- **Clustering Count:** For the clustering task, we randomly generate 2 to 10 clusters for each task.
Prompts: "You are a scatter plot visualization expert. Is there any cluster in this visualization? Can you tell me how many clusters are in this visualization?"
- **Outlier Count:** Different from the cluster recognition task, the scatter plot for the outlier detection will sample 1-5 outlier points without overlap in the visualization.
Prompts: "You are a scatter plot visualization expert. Is there any outlier in this visualization? Can you tell me how many outliers are in this visualization?"
- **Correlation Detection** For the correlation detection task, we randomly generated two scatter plot visualizations with different correlation efficient scores ranging from 0.1 to 1.0. It is worth noticing in the experiences that if the correlation in both images is very low (e.g., 0.1, 0.2) and LLM can not distinguish which one has a higher correlation but indicates both scatter plots have a low correlation, we will count this prediction correct.
Prompts: "You are a scatter plot visualization expert. which images have a high correlation?"

Parallel Coordinate

In parallel coordinate experiments, each task will be performed in 10 experiments, and the final results are aggregated as a percentage number (success rate) In each experiment, 500 points are generated randomly following a pre-defined pattern (e.g., number of clusters).

- **Clustering Count:** For the clustering task, we generate 1 to 10 clusters for each task.
Prompts: "You are a parallel coordinate visualization expert. Is there any cluster in this visualization? Can you tell me how many clusters are in this visualization?"
- **Outlier Count:** Different from the cluster recognition task, the outlier detection will sample 1-5 outlier points without overlap in the visualization.
Prompts: "You are a parallel coordinate visualization expert. Is there any outlier in this visualization? Can you tell me how many outliers are in this visualization?"
- **Correlation Detection** Compared with the scatter plot visualization, in the correlation task, we randomly select two attributions to be correlated and these two attributions are nearby in the parallel coordinate visualization.
Prompts: "You are a parallel coordinate visualization expert. Is there any correlation between these variables?"

Graph

It is worth noticing that we use the model to understand the result of an image without interaction operation. Therefore, we only use

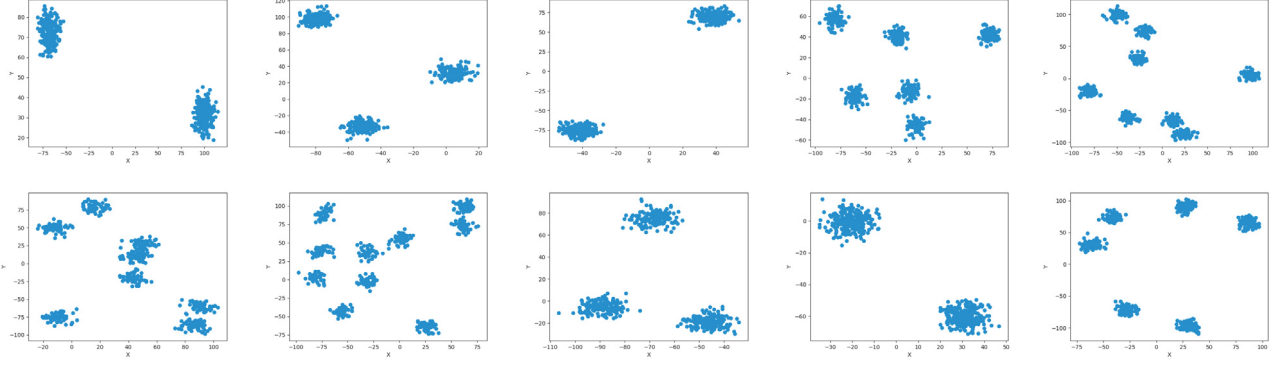


Figure 11: The randomly generated scatter plot with different numbers of clusters for LLM evaluation.

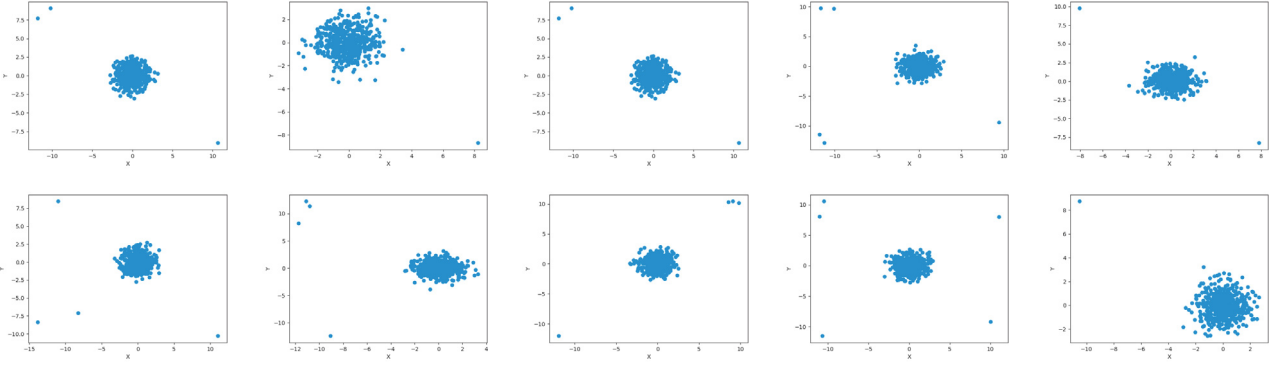


Figure 12: The randomly generated scatter plot with different numbers of outliers for LLM evaluation.

graph visualization which is visually interpretable (e.g., not edge or node clutter). We use a graph with 10 nodes and the overall sparsity is 20%. The final visualization is displayed by force-directed graph layout. Similarly, each experiment will be performed 10 times and each time the graph and connection are randomly generated.

Prompts used for the graph in LLM evaluation:

- **node count:** Prompts: "You are a graph visualization expert. How many nodes are in this visualization?"
- **find node:** Prompts: "You are a graph visualization expert. Is there a node named XXX in this visualization?"
- **connection:** Prompts: "You are a graph visualization expert. Is there a path from node XXX to node XXX?"
- **neighbor:** Prompts: "You are a graph visualization expert. What is the neighbor node of node XXX?"

Volume Rendering

Prompts used to evaluate the structure of interest recognition in the volume rendering:

- **Boston Teapot** "You are provided with several screenshots showing a volume rendering of the same CT data, for each image assess whether you can recognize the structure of interest, a teapot. Only assess for the structure of interest and not any other

structures you can recognize in the screenshot. Use only one of these options for assessment: 'Not recognizable', and 'Recognizable'. 'Not recognizable' means that the structure of interest cannot be identified in the image, even if another structure is recognizable. 'Recognizable' implies that both the structure of interest and its shape can be discerned in the screenshot.?

- **Visible Male** "You are provided with several screenshots showing a volume rendering of the same CT data, for each image assess whether you can recognize the structure of interest, a human face. Only assess for the structure of interest and not any other structures you can recognize in the screenshot. Use only one of these options for assessment: 'Not recognizable', and 'Recognizable'. 'Not recognizable' means that the structure of interest cannot be identified in the image, even if another structure is recognizable. 'Recognizable' implies that both the structure of interest and its shape can be discerned in the screenshot."

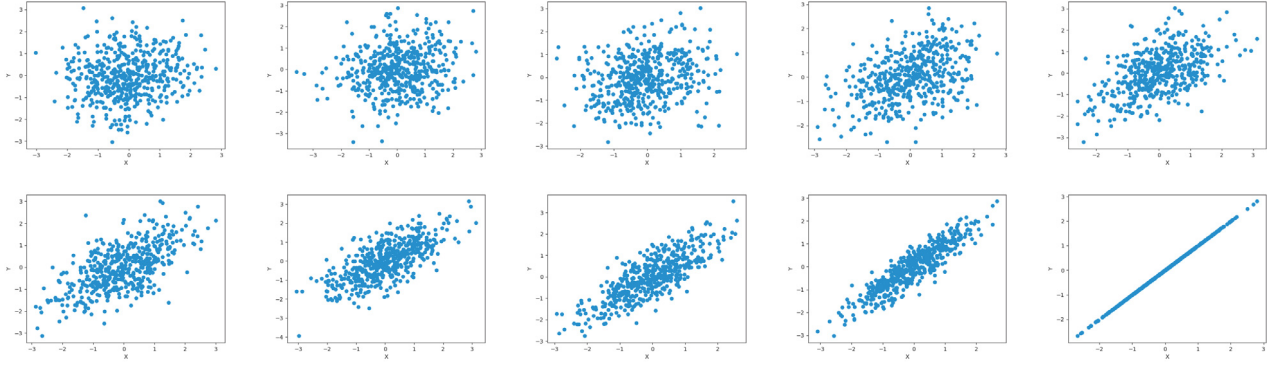


Figure 13: The randomly generated scatter plot with different correlation coefficients.

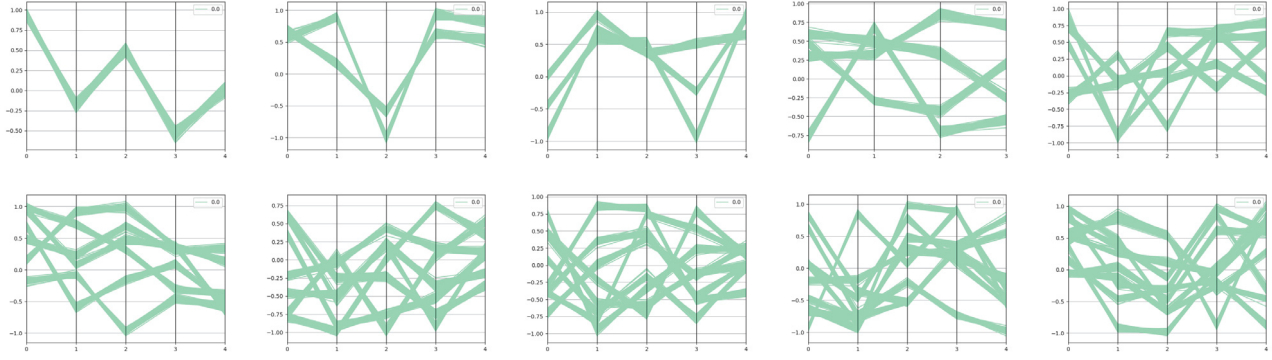


Figure 14: The randomly generated parallel coordinate with five dimensions and different numbers of clusters for LLM evaluation.

Algorithm 1 Opacity Transfer Function Adjustment - Used for the Heuristic-centric Action Plan

```

1:  $scalar\_range \leftarrow max\_val - min\_val$ 
2:  $window\_width \leftarrow scalar\_range / bins$ 
3:  $step\_size \leftarrow window\_width \times window\_factor$ 
4:  $start\_point \leftarrow min\_val$ 
5:  $end\_point \leftarrow start\_point + window\_width$ 
6: repeat
7:    $assessment\_result \leftarrow assess\_screenshot()$ 
8:   if  $assessment\_result == "not\ recognizable"$  then
9:      $start\_point \leftarrow start\_point + step\_size$ 
10:     $end\_point \leftarrow end\_point + step\_size$ 
11:   else if  $assessment\_result == "recognizable"$  then
12:      $start\_point \leftarrow start\_point + step\_size \times$ 
        $speed\_reduction$ 
13:      $end\_point \leftarrow end\_point + step\_size \times$ 
        $speed\_reduction$ 
14:   end if
15: until  $assessment\_result == "clear"$ 

```

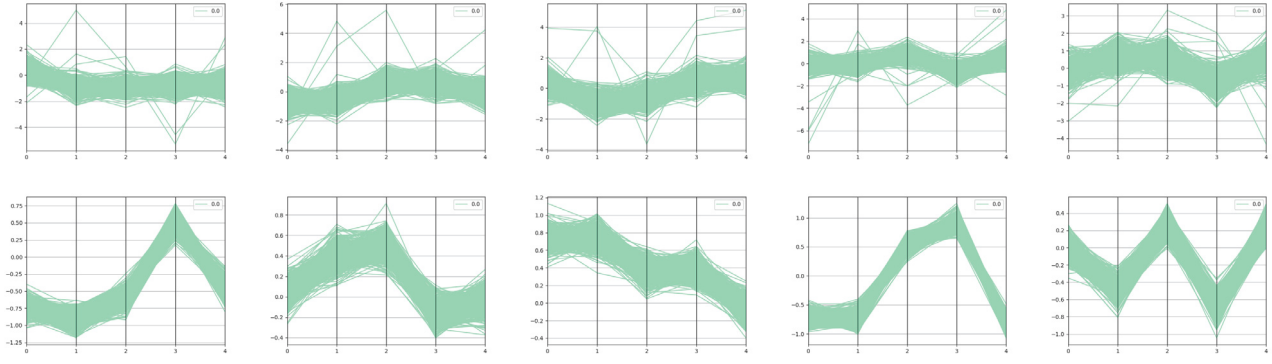


Figure 15: The randomly generated parallel coordinate with five dimensions and different numbers of outliers for LLM evaluation.

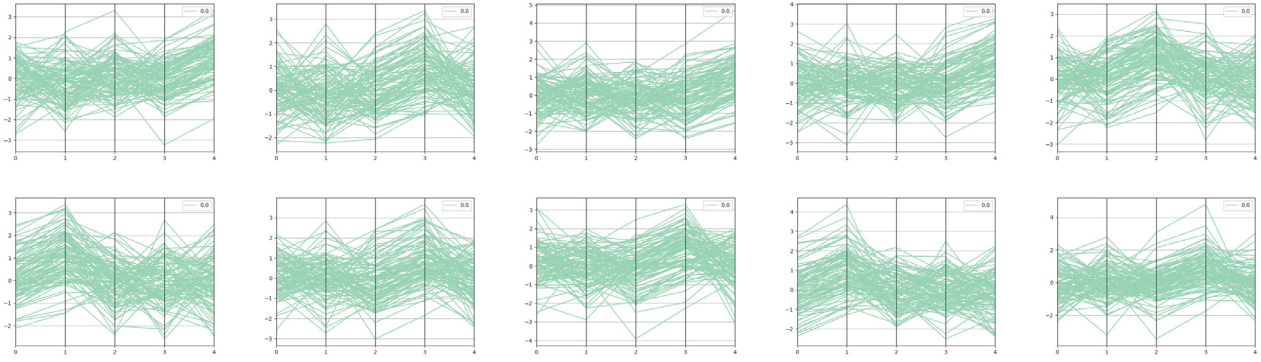


Figure 16: The randomly generated parallel coordinate with five dimensions and two of them have a high correlation.

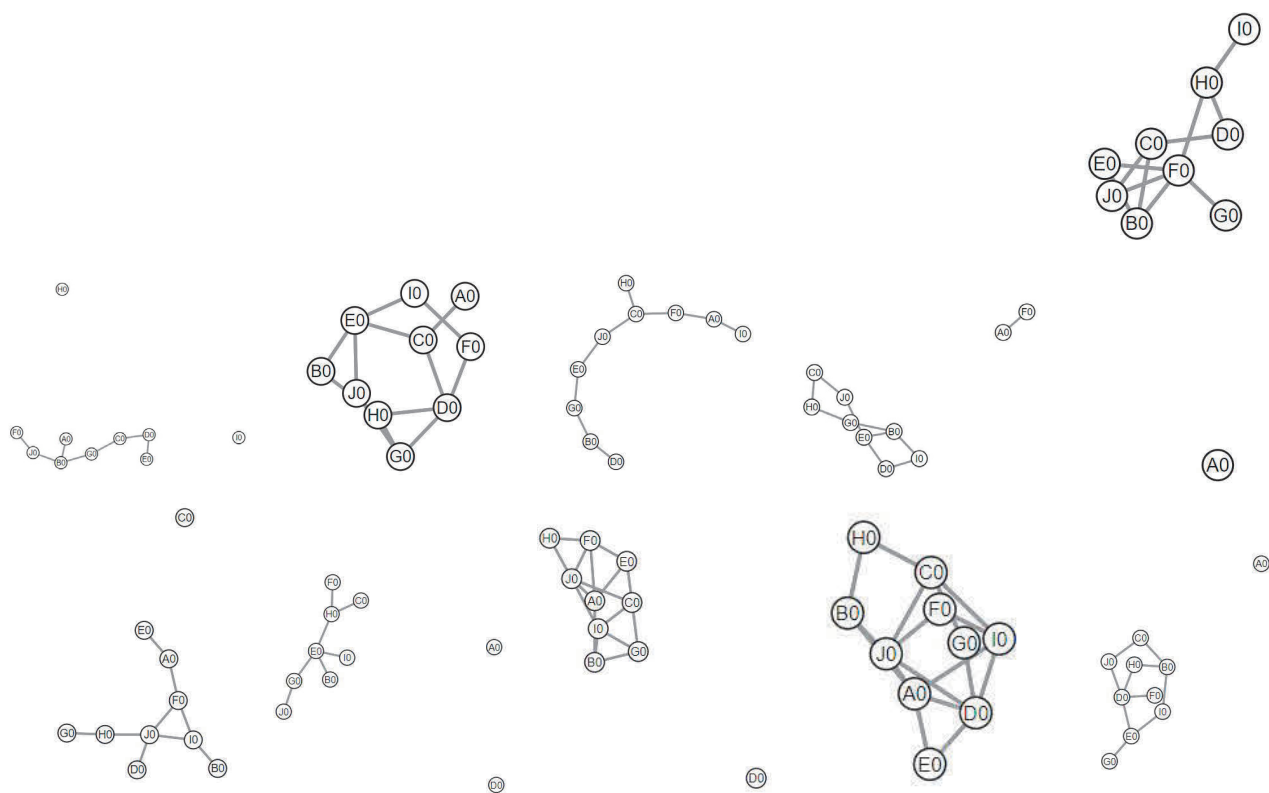


Figure 17: The randomly generated graph for graph exploration tasks

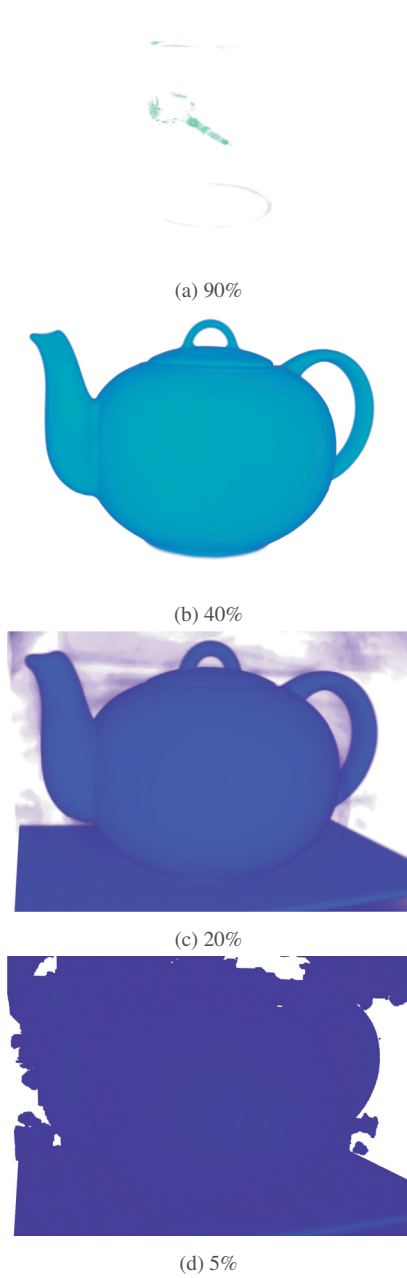


Figure 18: The Boston Teapot dataset volume rendered using the same color map but at varying opacity levels. Structure of interest: the teapot. The response from the LLM model was 18a: 'not recognizable', 18b: 'recognizable', 18c: 'recognizable', and 18d: 'not recognizable'

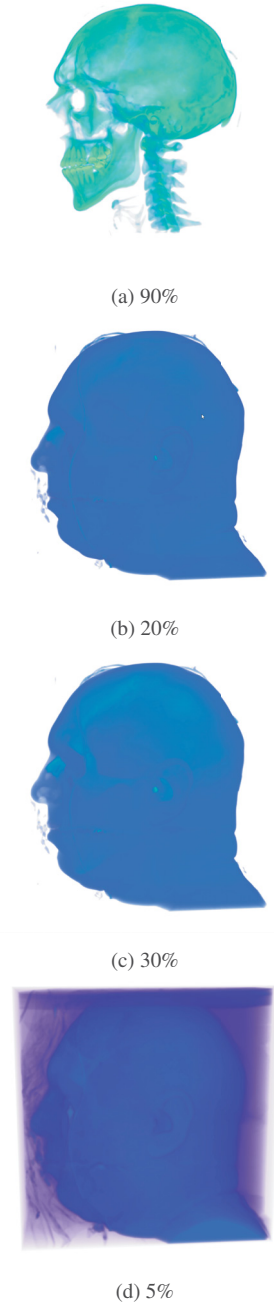


Figure 19: The Human Male dataset. In this figure, we conducted tests on the same set of screenshots, focusing on two distinct structures of interest: the male face and the bones. Notably, all images were accurately identified, except for the instance depicted in Figure 19d, where the presence of background partially occludes the head. The high degree of noise in this scenario appears to have affected the recognition of the skull. The agent's response for a 'male face' as the structure of interest: 19a: 'not recognizable', 19b: 'recognizable', 19c: 'recognizable', and 19d: 'not recognizable'. The agent's response for a 'bones' structure of interest: 19a: 'recognizable', 19b: 'not recognizable', 19c: 'not recognizable', and 19d: 'recognizable'.